

Potrzeba klienta

System do analizy częstości tekstu, która pozwoli określić, jakie słowa i tematy dominują w pliku tekstowym (tu przemówienia Bidena z 2021 i 2024), a także wskazać różnice w sposobie przekazu i priorytetach prezydenta w danym wystąpieniu. System ma wizualizować wyniki.

Etapy procesu tworzenia systemu informatycznego

1. Planowanie

Celem analizy jest **identyfikacja i porównanie najczęściej występujących słów w przemówieniu** (tu prezydenta Joe Bidena). Analiza pozwoli określić **dominujące tematy oraz potencjalne podobieństwa/różnice** dla każdego przemówienia.

2. Analiza (wymagań)

Proces obejmie **ekstrakcję tekstu, jego wstępne przetworzenie oraz analizę częstości występowania słów**, a także wizualizację wyników w formie **chmur słów oraz wykresów słupkowych**.

3. Projektowanie

Przygotowanie metod analizy tekstu i wizualizacji wyników

1. **Wczytanie tekstu przemówień** – import pliku tekstowego zawierającego przemówienie.
2. **Przetwarzanie tekstu** – oczyszczenie danych, usunięcie znaków interpunkcyjnych i konwersja tekstu do postaci tokenów.
3. **Usunięcie stop słów** – eliminacja słów o wysokiej częstości, ale niskiej wartości analitycznej (np. „i”, „oraz”, „dla”).
4. **Analiza częstości słów** – identyfikacja i porównanie najczęściej występujących terminów w obu przemówieniach.
5. **Wizualizacja wyników:**
 - **Wykresy słupkowe** – przedstawienie najczęściej używanych słów i ich liczebności.
 - **Chmury słów** – graficzne zobrazowanie częstości słów, gdzie większa czcionka oznacza wyższe występowanie.
 - **Porównanie wyników** – zestawienie najważniejszych różnic między przemówieniami pod kątem słownictwa i tematów.

4. Implementacja -wytworzenie kodu systemu

Wymagania funkcjonalne i нефункционалне dla skryptu w języku R, który analizuje dane tekstowe i tworzy chmurę słów.

Wymagania funkcjonalne:

1. Wczytywanie danych tekstowych:

System powinien umożliwiać użytkownikowi wybór i wczytanie pliku tekstowego z lokalnego dysku.

2. Analiza częstości występowania terminów:

System powinien analizować wczytany tekst w celu zidentyfikowania i zliczenia częstości występowania poszczególnych słów.

System powinien umożliwiać uwzględnienie lub pominięcie najczęściej występujących słów (tzw. stop words) podczas analizy.

3. Wizualizacja wyników w postaci chmury słów:

System powinien generować chmurę słów na podstawie wyników analizy częstości występowania terminów.

System powinien umożliwiać użytkownikowi dostosowanie parametrów chmury słów, takich jak:

- minimalna częstość występowania słowa, aby zostało uwzględnione w chmurze.
- maksymalna liczba słów wyświetlanych w chmurze.
- kolorystyka chmury słów, z możliwością wyboru różnych palet kolorów.

Wymagania нефункционалне:

1. Wydajność:

System powinien przetwarzać pliki tekstowe o wielkości do 10 MB w czasie nie dłuższym niż 30 sekund.

2. Użyteczność:

Interfejs użytkownika powinien być intuicyjny i umożliwiać łatwe wczytywanie plików.

3. Przenośność:

System powinien być kompatybilny z najnowszymi wersjami R oraz pakietów używanych do analizy tekstu i tworzenia chmur słów.

4. Bezpieczeństwo:

System powinien zapewniać, że wczytywane pliki są przetwarzane w sposób bezpieczny, bez możliwości wykonania niepożądanego kodu.